



UC San Diego

ECCV 2026 · MALMÖ

LangDriveCTRL: Natural Language Controllable Driving Scene Editing with Multi-modal Agents

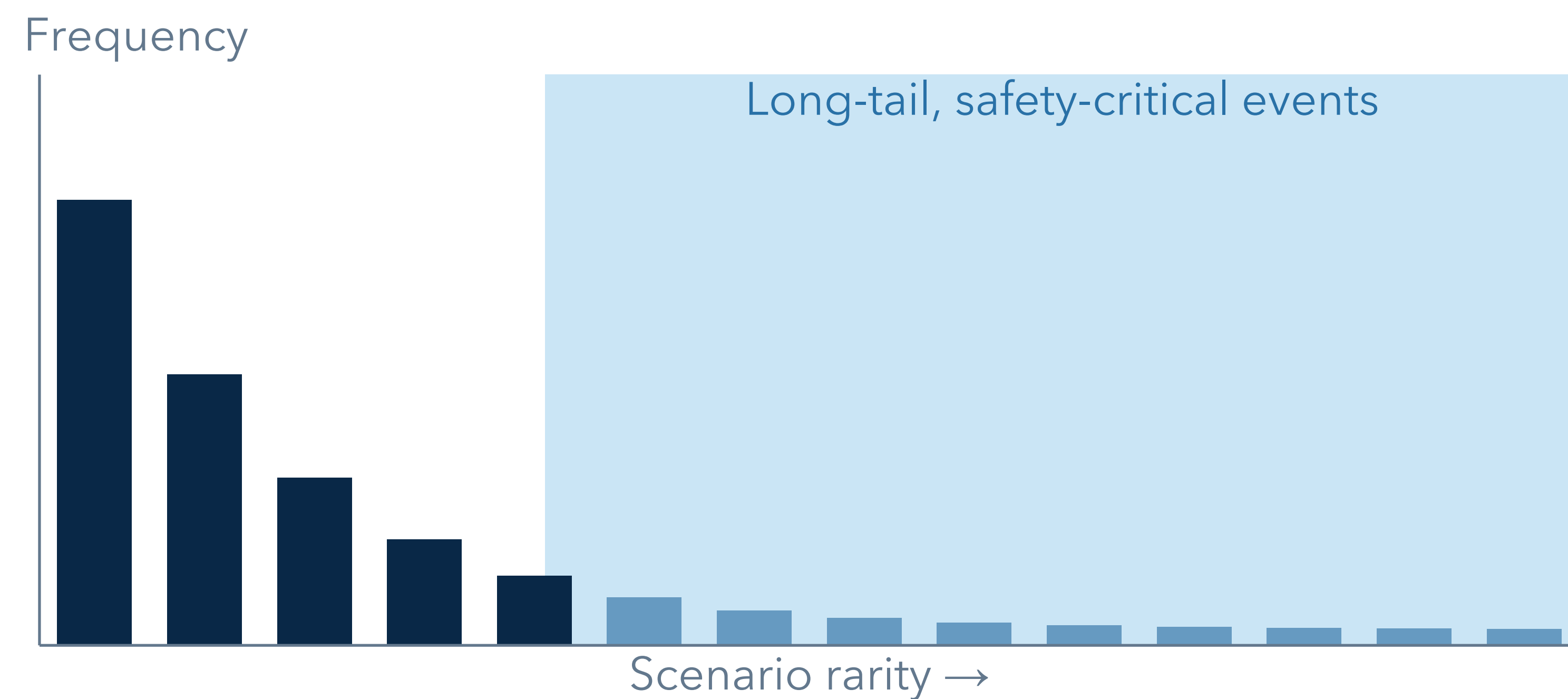
Yun He¹ Francesco Pittaluga² Ziyu Jiang²
Matthias Zwicker¹ Manmohan Chandraker^{2,3} Zaid Tasneem²

¹ University of Maryland, College Park ² NEC Labs America ³ UC San Diego



01 Long-Tail Scenario Generation for AV Stacks

Long-tail events dominate AV failures – yet are costly and rare to capture.



- Real long-tail data is **expensive & rare** to collect.
- Simulators (**CARLA**, **AirSim**) add diversity – but need hand-built 3D assets.
- They depend on **scripted scenarios** & manual human feedback.
- Scaling this pipeline is **expensive, slow and labor-intensive**.

02 Challenge of Driving Scene Editing

“Insert a blue sedan 3 meters to the left of ego vehicle, 7 meters ahead, and make it change to the right lane.”

Input					
Cosmos [World Model]					
ChatSim [Unimodal Agentic]					
Ours [Multi-modal Agentic]					

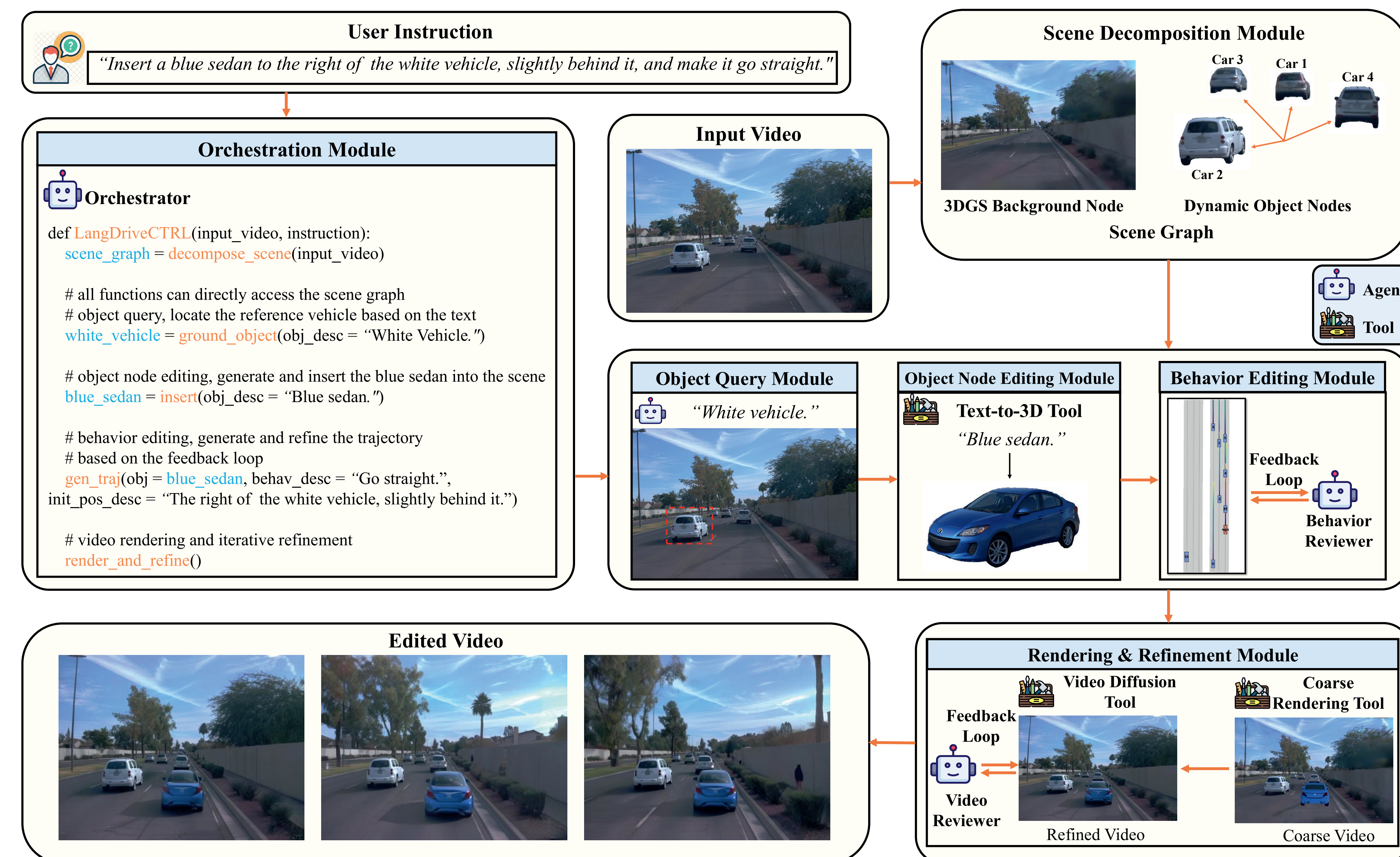
No prior method gets all four at once. **Cosmos** edits pixels, not 3D: no object-level control, and scene structure drifts. **ChatSim** reasons over text alone, mis-grounding targets, and never verifies its output, so errors accumulate.

03 Key Insights

Treat videos as explicit 3D scene graphs, not pixels.

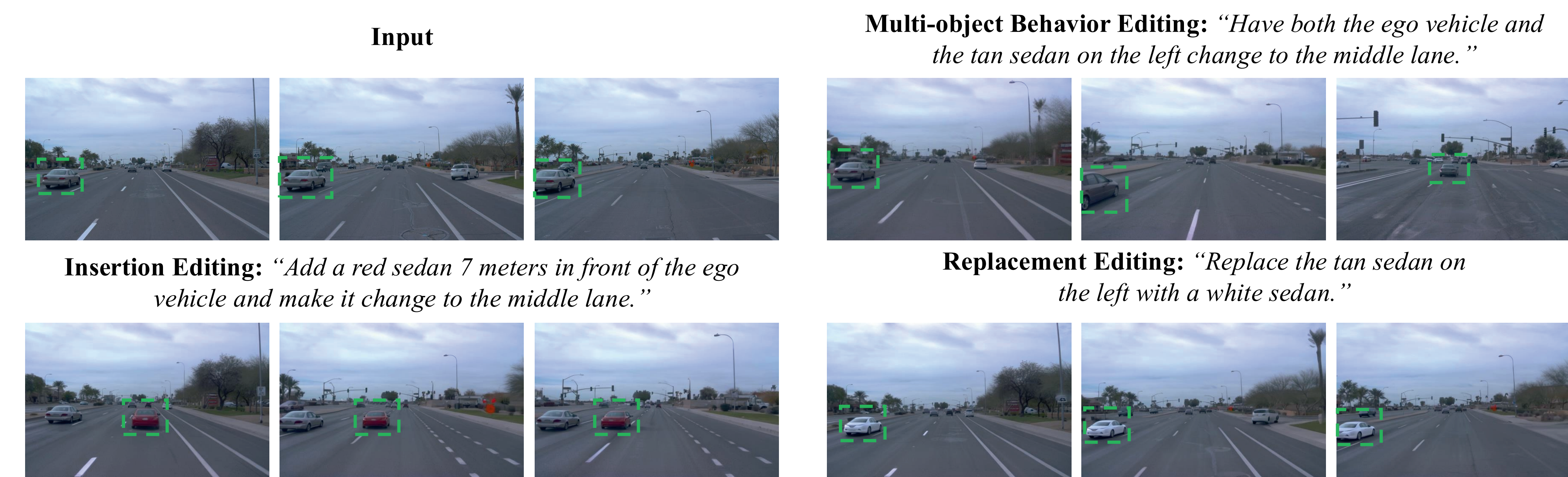
- Fine-grained control** needs **multi-modal reasoning** – grounding language jointly in visual appearance and traffic context.
- Photorealism + instruction alignment** need **feedback-driven iterative refinement** – behaviors and renderings reviewed and corrected in a closed loop.

04 Framework Overview



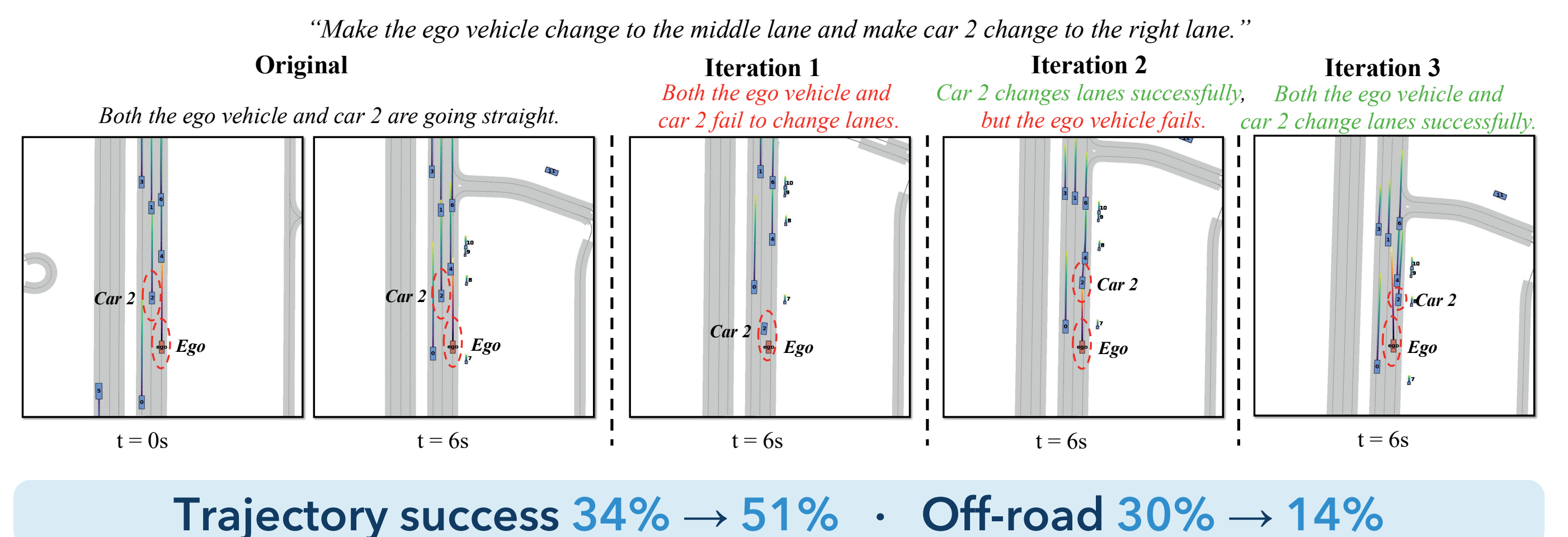
- Orchestrator** – decomposes the instruction into an executable agent/tool workflow.
- Object Grounding Agent** – matches free-form text to scene-graph nodes using appearance, motion and position.
- Behavior Editing Agent** – edits multi-object trajectories from language.
- Behavior & Video Reviewer Agents** – the Behavior Reviewer refines generated trajectories, the Video Reviewer refines the rendered result; both correct in a closed loop.

05 Controllable Editing Capabilities

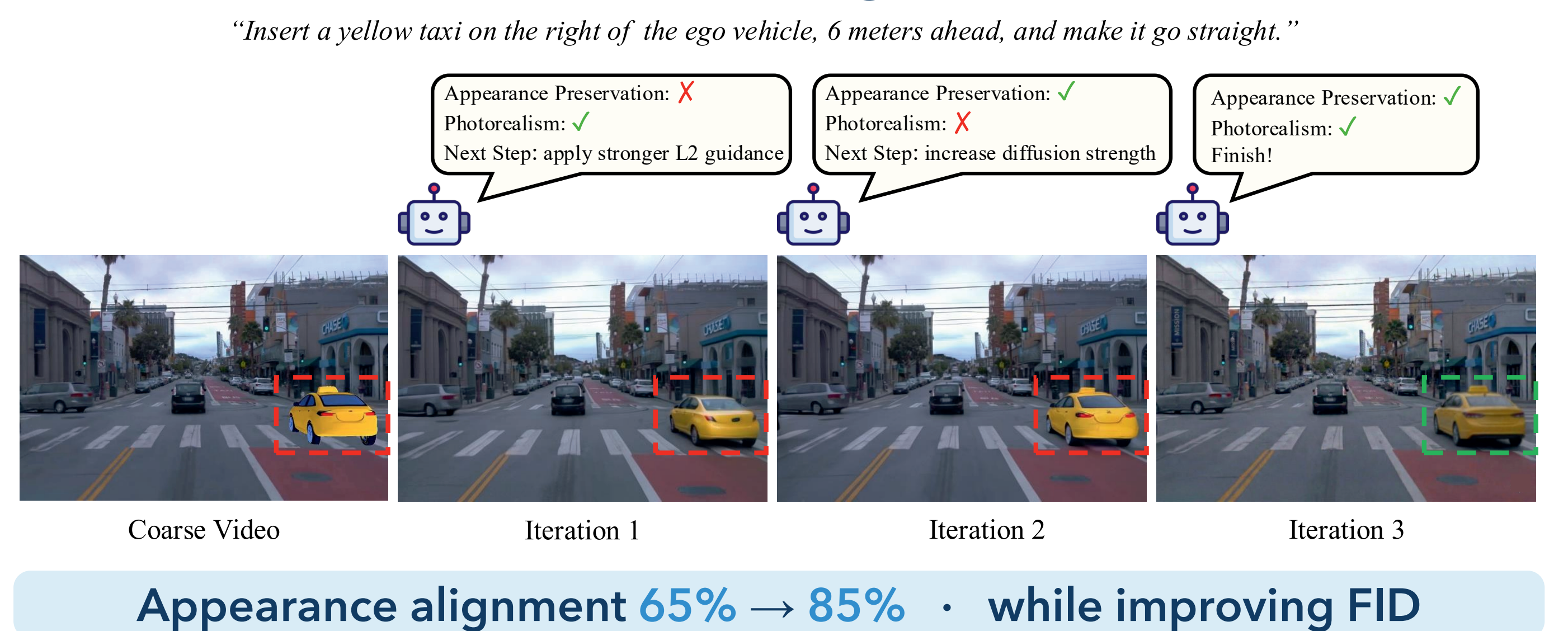


Insertion · Replacement · Removal · Multi-object behavior – every edit from one natural-language instruction.

06 Behavior Reviewer Agent



07 Video Reviewer Agent



08 vs. State of the Art



09 Quantitative Results

Human raters chose **Ours** as best far more often than Cosmos or ChatSim, at comparable runtime.

Method	Photoreal. FID / FVD ↓ user %	Instr. Align. App / Beh ↑ user %	Structure Str. ↓ user %	Traffic Col / Off ↓ user %	Time min ↓
Cosmos	33.4 / 798 34.6	46.3 / 32.9 10.5	74.1 19.1	3.2 / 4.0 35.5	16.9
ChatSim	47.7 / 606 4.8	42.3 / 26.6 3.9	46.5 7.4	27.6 / 24.7 5.6	19.3
Ours	32.9 / 467 54.2	82.2 / 71.7 60.4	34.6 65.0	0.6 / 1.7 48.3	17.1

App / Beh = appearance / behavior alignment (%). Col / Off = collision / off-road rate (%). "user %" = share of 26 raters choosing that method as best. Time on one A6000 GPU.